PSYCHOMETRIKA—VOL. 73, NO. 3, 487–502
SEPTEMBER 2008
DOI: 10.1007/S11336-008-9059-Y

ON SIMILARITY COEFFICIENTS FOR 2 × 2 TABLES AND CORRECTION FOR CHANCE

MATTHIJS J. WARRENS

LEIDEN UNIVERSITY

This paper studies correction for chance in coefficients that are linear functions of the observed proportion of agreement. The paper unifies and extends various results on correction for chance in the literature. A specific class of coefficients is used to illustrate the results derived in this paper. Coefficients in this class, e.g. the simple matching coefficient and the Dice/Sørenson coefficient, become equivalent after correction for chance, irrespective of what expectation is used. The coefficients become either Cohen's kappa, Scott's pi, Mak's rho, Goodman and Kruskal's lambda, or Hamann's eta, depending on what expectation is considered appropriate. Both a multicategorical generalization and a multivariate generalization are discussed.

Key words: indices of association, resemblance measures, correction for chance, Cohen's kappa, Scott's pi, Mak's rho, Goodman and Kruskal's lambda, Hamann's eta, simple matching coefficient, Dice/Sørenson coefficient.

1. Introduction

Measures of resemblance play an important role in many domains of data analysis. A similarity coefficient is a measure of association or agreement of two entities or variables. A well-known coefficient for two continuous variables is Pearson's product-moment correlation, but various other similarity coefficients may be used (see, e.g., Goodman & Kruskal, 1954; Zegers & Ten Berge, 1985; Gower & Legendre, 1986). In this paper we focus on similarity coefficients that can be defined using the four dependent proportions, a, b, c, and d, presented in Table 1. Instead of probabilities, Table 1 may also be defined on counts or frequencies; probabilities are used here for notational convenience.

The data in Table 1 may be obtained from a 2×2 reliability study: a, b, c, and d are observed proportions resulting from classifying m persons using a dichotomous response (Fleiss, 1975; Bloch & Kraemer, 1989; Blackman & Koval, 1993). In cluster analysis, Table 1 may be the result of comparing partitions from two clustering methods: a is the proportion of object pairs that were placed in the same cluster according to both clustering methods, b (c) is the proportion of pairs that were placed in the same cluster according to one method but not according to the other, and d is the proportion of pairs that were not in the same cluster according to either of the methods (Albatineh, Niewiadomska-Bugaj & Mihalko, 2006; Steinley, 2004).

Numerous 2×2 resemblance measures have been proposed in the literature (Gower & Legendre, 1986; Krippendorff, 1987; Hubálek, 1982; Baulieu, 1989; and Albatineh et al., 2006). Let a similarity coefficient be denoted by S. Table 2 presents ten similarity coefficients that will be used to illustrate the results in this paper. Following Sokal and Sneath (1963, p. 128) and Albatineh et al. (2006), the convention is adopted of calling a coefficient by its originator or the first

The author thanks two anonymous reviewers for their helpful comments and valuable suggestions on earlier versions of this article.

Requests for reprints should be sent to Matthijs J. Warrens, Psychometrics and Research Methodology Group, Leiden University Institute for Psychological Research, Leiden University, Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, The Netherlands. E-mail: warrens@fsw.leidenuniv.nl

TABLE 1.
Bivariate proportions table for binary variables.

		Variable two		
	Proportions	Value 1	Value 2	Total
Variable one	Value 1	a	b	p_1
	Value 2	c	d	q_1
	Total	p_2	q_2	1

we know to propose it. The coefficients in Table 2 may be considered both as population parameters as well as sample statistics; in this paper we use the latter. Some of these coefficients have been proposed in different domains of data analysis, but turn out to be equivalent after recoding.

If the two variables are statistically independent, we may desire that the theoretical value of a similarity coefficient be zero. Coefficient S_{Cohen} satisfies this requirement; coefficients S_{SM} and S_{Cze} do not. If a coefficient does not have zero value under statistical independence, it may be corrected for agreement due to chance (Fleiss, 1975; Zegers, 1986; Krippendorff, 1987; Albatineh et al., 2006). After correction for chance, a similarity coefficient S has a form

$$CS = \frac{S - E(S)}{1 - E(S)}, \tag{1}$$

where expectation $E(S)$ is conditional upon fixed marginal proportions in Table 1. Various authors have noted that some coefficients become equivalent after correction (1). For example, Fleiss (1975) and Zegers (1986) showed that S_{SM} and S_{Cze} become S_{Cohen} after correction (1). In addition, Zegers (1986) showed that S_{Ham}, and Fleiss (1975) showed that S_{GK1} and S_{RG}, become S_{Cohen} after correction for chance.

Albatineh et al. (2006) studied correction (1) for a specific family of coefficients. They showed that coefficients may coincide after correction for chance, irrespective of what expectation is used. The main result of their paper is Proposition 1 in Section 3. In this paper, we continue the general approach by Albatineh et al. (2006) and present several new results with respect to correction (1).

The paper is organized as follows. Similar to Albatineh et al. (2006) correction (1) is studied for a general family of coefficients. This family of coefficients, of a form $S = \lambda + \mu(a + d)$, is introduced in the next section. Section 3 is used to present the main results. In addition to a powerful result by Albatineh et al. (2006), Section 3 considers two additional functions. If coefficients are related by one of these functions, they become equivalent after correction (1), irrespective of what expectation $E(S)$ is used.

Additional results may be obtained by considering different expectations $E(S)$. The specific results in Section 4 unify and extend the findings for individual coefficients in Fleiss (1975) and Zegers (1986). Section 5 discusses corrected coefficients and some of their properties. Also in Section 5, we discuss a generalization of an inequality in Blackman and Koval (1993) for Cohen's kappa and Scott's pi. Sections 6 and 7 discuss two natural generalizations of the results in Sections 3 to 5. Section 6 presents a multicategorical extension; Section 7 describes a family of multivariate coefficients. Section 8 contains the discussion.

2. A Family of Coefficients

Consider a family $\mathcal{L}$ of coefficients of a form $S = \lambda + \mu(a + d)$, where proportions a and d are defined in Table 1, and where λ and μ, different for each coefficient, depend on the marginal

TABLE 2.
Ten 2×2 similarity coefficients.

Symbol	Formula	Source
S_{SM}	$a + d$	Sokal and Michener (1958), Rand (1971), Brennan and Light (1974)
S_{Ham}	$a - b - c + d$	Hamann (1961), Hubert (1977)
S_{Cze}	$\frac{2a}{p_1+p_2}$	Czekanowski (1932), Dice (1945), Sørenson (1948), Nei and Li (1979)
S_{GK1}	$\frac{2a-b-c}{2a+b+c}$	Goodman and Kruskal (1954)
S_{GK2}	$\frac{2d-b-c}{b+c+2d}$	
S_{GK3}	$\frac{2\min(a,d)-b-c}{2\min(a,d)+b+c}$	
S_{NS}	$\frac{2d}{q_1+q_2}$	No source
S_{RG}	$\frac{a}{p_1+p_2} + \frac{d}{q_1+q_2}$	Rogot and Goldberg (1966)
S_{Scott}	$\frac{4ad-(b+c)^2}{(p_1+p_2)(q_1+q_2)}$	Scott (1955)
S_{Cohen}	$\frac{2(ad-bc)}{p_1q_2+p_2q_1}$	Cohen (1960)

probabilities of Table 1. Since $S_{\text{SM}} = a + d$, all members in $\mathcal{L}$ family are linear transformations of S_{SM}, the observed proportion of agreement, given the marginal probabilities. Clearly, S_{SM} is in $\mathcal{L}$ family. Furthermore, all ten coefficients in Table 2 are in $\mathcal{L}$ family.

Example 1. Coefficient S_{Cze} was independently proposed by Czekanowski (1932), Dice (1945), and Sørenson (1948). The coefficient is often attributed to Dice (1945), and it was also derived by Nei and Li (1979). Bray (1956) noted that coefficient S_{Cze} could be found in Gleason (1920). Coefficient S_{Cze}

$$S_{\text{Cze}} = \frac{2a}{p_1 + p_2} = \frac{(a + d) - 1}{p_1 + p_2} + 1.$$

Thus, coefficient S_{Cze} can be written in a form $S_{\text{Cze}} = \lambda + \mu(a + d)$, where

$$\lambda = \frac{-1}{p_1 + p_2} + 1 \quad \text{and} \quad \mu = \frac{1}{p_1 + p_2}.$$

Example 2. Scott (1955) proposed a measure of interrater-reliability denoted by the symbol pi. For two dichotomized variables Scott's pi

$$S_{\text{Scott}} = \frac{4ad - (b + c)^2}{(p_1 + p_2)(q_1 + q_2)}.$$

With respect to the numerator of S_{Scott}, we have

$$a(1 - a - b - c) - \frac{(b + c)^2}{4} = a - \frac{(a + b)^2}{4} - \frac{(a + c)^2}{4} - \frac{(a + b)(a + c)}{2} = a - \left(\frac{p_1 + p_2}{2}\right)^2.$$

Similarly we have

$$d(1 - b - c - d) - \frac{(b + c)^2}{4} = d - \left(\frac{q_1 + q_2}{2}\right)^2.$$

Thus, coefficient S_{Scott}

$$S_{\text{Scott}} = \frac{4(a+d) - (p_1+p_2)^2 - (q_1+q_2)^2}{2(p_1+p_2)(q_1+q_2)}$$

can be written in a form $S_{\text{Scott}} = \lambda + \mu(a+d)$ where

$$\lambda = \frac{-(p_1+p_2)^2 - (q_1+q_2)^2}{2(p_1+p_2)(q_1+q_2)} \quad \text{and} \quad \mu = \frac{2}{(p_1+p_2)(q_1+q_2)}.$$

Example 3. The best-known index for interrater-reliability is the kappa-statistic proposed by Cohen (1960). Since

$$ad - bc = a(1-a-b-c) - bc = a - (a+b)(a+c) = a - p_1p_2 \quad \text{and}$$
$$ad - bc = d(1-b-c-d) - bc = d - (b+d)(c+d) = d - q_1q_2,$$

Cohen's kappa for two dichotomized variables is given by

$$S_{\text{Cohen}} = \frac{2(ad-bc)}{p_1q_2 + p_2q_1} = \frac{(a+d) - p_1p_2 - q_1q_2}{p_1q_2 + p_2q_1}.$$

Coefficient S_{Cohen} can be written in a form $S_{\text{Cohen}} = \lambda + \mu(a+d)$, where

$$\lambda = \frac{-p_1p_2 - q_1q_2}{p_1q_2 + p_2q_1} \quad \text{and} \quad \mu = \frac{1}{p_1q_2 + p_2q_1}.$$

Since $a = p_2 - q_1 + d$, probabilities a and d are also linear in $(a+d)$. Linear in $(a+d)$ is therefore equivalent to linear in a and linear in d. Furthermore, Albatineh et al. (2006) studied coefficients that are linear in $\sum\sum n_{ij}^2$, where n_{ij} is the number of data points placed in cluster i according to the first clustering method and in cluster j according to the second clustering method. Because $ma = (\sum\sum n_{ij}^2 - m)/2$, linear in $\sum\sum n_{ij}^2$ is equivalent to linear in a and equivalent to linear in $(a+d)$.

A well-known similarity measure that cannot be written in a form $S = \lambda + \mu(a+d)$ is coefficient

$$S_{\text{Jac}} = \frac{a}{a+b+c} = \frac{a}{p_1+p_2-a}$$

by Jaccard (1912). Other examples of coefficients that do not belong to $\mathcal{L}$ family can be found in Albatineh et al. (2006) and Baulieu (1989).

3. Main Results

Albatineh et al. (2006) showed that correction (1) is relatively simple for coefficients that belong to $\mathcal{L}$ family. Two members in $\mathcal{L}$ family become equivalent after correction for chance agreement if they have the same ratio (2).

Proposition 1 (Albatineh et al., 2006, p. 309). *Two members in $\mathcal{L}$ family become identical after correction* (1) *if they have the same ratio*

$$\frac{1-\lambda}{\mu}. \tag{2}$$

Proof: $E(S) = E[\lambda + \mu(a+d)] = \lambda + \mu E(a+d)$ and consequently the CS becomes

$$CS = \frac{S - E(S)}{1 - E(S)} = \frac{\lambda + \mu(a+d) - \lambda - \mu E(a+d)}{1 - \lambda - \mu E(a+d)}$$
$$= \frac{a + d - E(a+d)}{\mu^{-1}(1-\lambda) - E(a+d)}. \qquad \square \qquad (3)$$

Thus, the value of a similarity coefficient after correction for chance depends on ratio (2), where λ and μ characterize the particular measure within $\mathcal{L}$ family.

Corollary 1 below extends Corollary 4.2(i) in Albatineh et al. (2006) from three measures (S_{SM}, S_{Ham}, and S_{Cze}) to the ten coefficients in Table 2. The coefficients in Table 2 coincide after correction (1), irrespective of what expectation $E(S)$ is used.

Corollary 1. *Coefficients* S_{SM}, S_{Ham}, S_{Cze}, S_{GK1}, S_{GK2}, S_{GK3}, S_{NS}, S_{RG}, S_{Scott}, *and* S_{Cohen} *become equivalent after correction* (1).

Proof: By Proposition 1 it suffices to inspect ratio (2). Using the formulas of λ and μ corresponding to each coefficient we obtain the ratio (2)

$$\frac{1-\lambda}{\mu} = 1 \qquad (4)$$

for all ten coefficients. Only the proofs for coefficients S_{Scott} and S_{Cohen} are presented. Using the formulas for λ and μ from Example 2 we obtain the ratio (2)

$$\frac{1-\lambda}{\mu} = \frac{2(p_1+p_2)(q_1+q_2) + (p_1+p_2)^2 + (q_1+q_2)^2}{4} = \frac{(p_1+p_2+q_1+q_2)^2}{4} = 1.$$

Using the formulas for λ and μ from Example 3 we obtain the ratio (2)

$$\frac{1-\lambda}{\mu} = p_1q_2 + p_2q_1 + p_1p_2 + q_1q_2 = (p_1+q_1)(p_2+q_2) = 1. \qquad \square$$

Note that $(1-\lambda)/\mu = 1$ for all coefficients in Table 2 (ratio (4)). The value 1 is also the maximum value regardless of the marginal probabilities of these similarity coefficients.

Due to Proposition 1, ratio (2) may be used to inspect whether coefficients become equivalent after correction for chance. Alternatively, it can be shown that coefficients that have a specific relationship coincide after correction. In the remainder of this section we consider two functions that may relate similarity coefficients:

$$S_2 = 2S_1 - 1 \quad \text{and} \quad S_3 = \frac{S_1 + S_2}{2}.$$

Both functions may be used to construct new resemblance measures from existing similarity coefficients. It is not difficult to show that $S_2 = 2S_1 - 1$ is in $\mathcal{L}$ family if and only if S_1 is in $\mathcal{L}$ family, and if S_1 and S_2 are in $\mathcal{L}$ family, then $S_3 = (S_1 + S_2)/2$ is in $\mathcal{L}$ family. Two coefficients S_1 and S_2 that are related by $S_2 = 2S_1 - 1$ become equivalent after correction for chance.

Proposition 2. *Let* S_1 *be a member in* $\mathcal{L}$ *family.* S_1 *and* $S_2 = 2S_1 - 1$ *become identical after correction* (1).

Proof: $S_2 = 2\lambda + 2\mu(a+d) - 1$ and $E(S_2) = 2\lambda - 1 + 2\mu E(a+d)$. Consequently the CS_2 becomes

$$CS_2 = \frac{2\lambda + 2\mu(a+d) - 1 - 2\lambda - 2\mu E(a+d) + 1}{1 - 2\lambda - 2\mu E(a+d) + 1} = \frac{\lambda + \mu(a+d) - \lambda - \mu E(a+d)}{1 - \lambda - \mu E(a+d)}$$
$$= \frac{S_1 - E(S_1)}{1 - E(S_1)} = CS_1. \qquad \square$$

Example 4. Various similarity coefficients have a relationship $S_2 = 2S_1 - 1$. Examples from Table 2 are $S_{\text{Ham}} = 2S_{\text{SM}} - 1$, $S_{\text{GK1}} = 2S_{\text{Cze}} - 1$, and $S_{\text{GK2}} = 2S_{\text{NS}} - 1$. Due to either Proposition 1 with Corollary 1 or Proposition 2, these coefficients coincide after correction (1).

Theorem 1. *Let S_i for $i = 1, 2, \ldots, n$ be members in $\mathcal{L}$ family that become identical after correction* (1). *Then S_i for $i = 1, 2, \ldots, n$ and the arithmetic mean*

$$AM = \frac{1}{n}\sum_{i=1}^{n} S_i \tag{5}$$

become equivalent after correction (1).

Remark. The original proof has been simplified with the help of an anonymous referee.

Proof:

$$E(AM) = \frac{1}{n}\left(\sum_{i=1}^{n}\lambda_i + \sum_{i=1}^{n}\mu_i E(a+d)\right). \tag{6}$$

Using (5) and (6) in (1) we obtain

$$CS = \frac{a+d-E(a+d)}{y - E(a+d)} \quad \text{where} \quad y = \frac{n - \sum_{i=1}^{n}\lambda_i}{\sum_{i=1}^{n}\mu_i}.$$

Let

$$x = \frac{1-\lambda_1}{\mu_1} = \frac{1-\lambda_2}{\mu_2} = \cdots = \frac{1-\lambda_n}{\mu_n}.$$

Due to Proposition 1, it must be shown that ratio y equals ratio x. We have

$$y = \frac{\sum_{i=1}^{n}(1-\lambda_i)}{\sum_{i=1}^{n}\mu_i} = \frac{\sum_{i=1}^{n} x\mu_i}{\sum_{i=1}^{n}\mu_i} = \frac{x\sum_{i=1}^{n}\mu_i}{\sum_{i=1}^{n}\mu_i} = x.$$

This completes the proof. $\square$

Example 5. Coefficient

$$S_{\text{RG}} = \frac{a}{2a+b+c} + \frac{d}{b+c+2d} = \frac{S_{\text{Cze}} + S_{\text{NS}}}{2}$$

is the arithmetic mean of S_{Cze} and S_{NS}. Due to either Proposition 1 with Corollary 1 or Theorem 1, these three coefficients become equivalent after correction (1).

4. Specific Results

Remember that (4) holds for all coefficients in Table 2. Due to Corollary 1 these coefficients coincide after correction (1). The corrected coefficient corresponding to the resemblance measures in Corollary 1 has a form

$$\frac{a+d-E(a+d)}{1-E(a+d)}. \tag{7}$$

Coefficient (7) may be obtained by using (4) in (3). Since expectation $E(a+d)$ is unspecified, coefficient (7) is a general corrected coefficient. Specific cases of (7) may be obtained by specifying $E(a+d)$ in (7).

Different opinions have been stated on what the appropriate expectations are for the 2×2 contingency table. Detailed discussions on the various ways of regarding data as the product of chance can be found in Krippendorff (1987), Mak (1988), Bloch and Kraemer (1989), and Pearson (1947). In cluster analysis it is general consensus that the popular coefficient S_{SM}, called the Rand index, should be corrected for agreement due to chance (Morey & Agresti, 1984; Hubert & Arabie, 1985), although there is some debate on what expectation is appropriate (Hubert & Arabie, 1985; Steinley, 2004; Albatineh et al., 2006). We consider five examples of $E(a+d)$.

Example 6a. Suppose it is assumed that the frequency distribution underlying the two variables in Table 1 is the same for both variables (Scott, 1955; Krippendorff, 1987, p. 113). Coefficients used in this case are sometimes referred to as agreement indices. The common parameter p must be either known or it must be estimated from p_1 and p_2. Different functions may be used. For example, Scott (1955) and Krippendorff (1987) used the arithmetic mean

$$p = \frac{p_1 + p_2}{2}.$$

Following Scott (1955) and Krippendorff (1987, p. 113) we have

$$E(a+d)_{\text{Scott}} = \left(\frac{p_1+p_2}{2}\right)^2 + \left(\frac{q_1+q_2}{2}\right)^2.$$

Let m denote the number of elements of the binary variables. Mak (1988) proposed the expectation

$$E(a+d)_{\text{Mak}} = 1 - \frac{m(p_1+p_2)(q_1+q_2)-(b+c)}{2(m-1)}$$

(see also, Blackman & Koval, 1993).

Example 6b. Instead of a single distribution function, it may be assumed that the data are a product of chance concerning two different frequency distributions, each with its own parameter (Cohen, 1960; Krippendorff, 1987). Coefficients used in this case are sometimes referred to as association indices. The expectation of an entry in Table 1 under statistical independence is defined by the product of the marginal probabilities. We have

$$E(a+d)_{\text{Cohen}} = p_1 p_2 + q_1 q_2.$$

The expectation $E(a+d)_{\text{Cohen}}$ can be obtained by considering all permutations of the observations of one of the two variables, while preserving the order of the observations of the other variable. For each permutation the value of $(a+d)$ can be determined. The arithmetic mean of these values is $p_1 p_2 + q_1 q_2$.

Example 6c. A third possibility is that there are no relevant underlying continua. For this case two forms of $E(a+d)$ may be found in the literature. Note that a and d in Table 1 may be interpreted as the proportions of positive and negative matches, whereas b and c are the proportions of nonmatching observations. Goodman and Kruskal (1954, p. 757) used expectation

$$E(a+d)_{\mathrm{GK}} = \frac{\max(p_1+p_2, q_1+q_2)}{2} = \frac{2\max(a,d)+b+c}{2}.$$

Expectation $E(a+d)_{\mathrm{GK}}$ focuses on the largest group of matching observations. According to Krippendorff (1987, p. 114) an equity coefficient is characterized by expectation

$$E(a+d)_{\mathrm{Kripp}} = \frac{1}{2}.$$

In the case of association (Example 6b) the observations are regarded as ordered pairs. In the case of agreement (Example 6a) the observations are considered as pairs without regard for their order; a mismatch is a mismatch regardless of the kind. In the case of equity one only distinguishes between matching and nonmatching observations (cf. Krippendorff, 1987).

Theorem 2 below unifies and extends findings in Fleiss (1975) and Zegers (1986) on what coefficients become Cohen's kappa after correction for chance. Depending on what expectation $E(a+d)$ from Examples 6a to 6c is used, the coefficients in Table 2 become, after correction for chance, either Scott's (1955) pi (S_{Scott}), Cohen's (1960) kappa (S_{Cohen}), Goodman and Kruskal's (1954) lambda (S_{GK3}), Hamann's (1961) eta (S_{Ham}), or Mak's (1988) rho. The latter coefficient can be written as

$$S_{\mathrm{Mak}} = \frac{4mad - m(b+c)^2 + (b+c)}{m(p_1+p_2)(q_1+q_2) - (b+c)}$$

where m is the length of the binary variables.

Theorem 2. *Let S be a member in $\mathcal{L}$ family for which ratio (4) holds. If the appropriate expectation is*

(i)	$E(a+d)_{\mathrm{Scott}}$,	*then S becomes* S_{Scott},
(ii)	$E(a+d)_{\mathrm{Mak}}$,	*then S becomes* S_{Mak},
(iii)	$E(a+d)_{\mathrm{Cohen}}$,	*then S becomes* S_{Cohen},
(iv)	$E(a+d)_{\mathrm{GK}}$,	*then S becomes* S_{GK3},
(v)	$E(a+d)_{\mathrm{Kripp}}$,	*then S becomes* S_{Ham},

after correction (1).

Proof: (i): Using $E(a+d)_{\mathrm{Scott}}$ in (7) we obtain an index of which the numerator equals

$$a+d-\left(\frac{p_1+p_2}{2}\right)^2-\left(\frac{q_1+q_2}{2}\right)^2 = 2ad - \frac{(b+c)^2}{2} \tag{8}$$

(see Example 2) and the denominator equals

$$\frac{(p_1+p_2+q_1+q_2)^2-(p_1+p_2)^2-(q_1+q_2)^2}{4} = \frac{(p_1+p_2)(q_1+q_2)}{2}. \tag{9}$$

Dividing the right-hand part of (8) by the right-hand part of (9) we obtain

$$\frac{4ad-(b+c)^2}{(p_1+p_2)(q_1+q_2)} = S_{\text{Scott}}.$$

(ii): Using $E(a+d)_{\text{Mak}}$ in (7) and multiplying the result by $2(m-1)$ we obtain an index of which the numerator equals

$$\begin{aligned} &2(a+d-1)(m-1)+m(p_1+p_2)(q_1+q_2)-(b+c)\\ &\quad = m(2a+b+c)(b+c+2d)-2m(b+c)+(b+c), \end{aligned} \tag{10}$$

and the denominator equals

$$m(p_1+p_2)(q_1+q_2)-(b+c). \tag{11}$$

We have

$$\begin{aligned} &(2a+b+c)(b+c+2d)-2(b+c)\\ &\quad = 4ad+(2a+2d)(b+c)+(b+c)^2-2(b+c)\\ &\quad = 4ad+(2a+2d-2)(b+c)+(b+c)^2\\ &\quad = 4ad-2(b+c)^2+(b+c)^2 = 4ad-(b+c)^2. \end{aligned} \tag{12}$$

Using the right-hand part of (12), numerator (10) can be written as

$$m\big[4ad-(b+c)^2\big]+(b+c). \tag{13}$$

Dividing (13) by (11) we obtain coefficient S_{Mak}.

(iii): Using $E(a+d)_{\text{Cohen}}$ in (7) we obtain

$$\frac{a+d-p_1p_2-q_1q_2}{(p_1+q_1)(p_2+q_2)-p_1p_2-q_1q_2} = \frac{2(ad-bc)}{p_1q_2+p_2q_1} = S_{\text{Cohen}}.$$

(iv): Using $E(a+d)_{\text{GK}}$ in (7) we obtain

$$\frac{2[a+d-\max(a,d)]-b-c}{2-2\max(a,d)-b-c} = \frac{2\min(a,d)-b-c}{2\min(a,d)+b+c} = S_{\text{GK3}}.$$

(v): Using $E(a+d)_{\text{Kripp}}$ in (7) we obtain

$$2(a+d)-1 = a-b-c+d = S_{\text{Ham}}. \qquad \square$$

5. Corrected Coefficients

The coefficients in Table 2 become either S_{Scott}, S_{Mak}, S_{Cohen}, S_{GK3}, or S_{Ham}, depending on what expectation $E(a+d)$ is used. Note that corrected coefficients S_{Scott}, S_{Cohen}, S_{GK3}, and S_{Ham} belong to the class of resemblance measures that is considered in Corollary 1 and Theorem 2. This suggests that corrected coefficients may have some interesting properties, which are the topic of this section. If $E(S)$ in (1) depends on the marginal probabilities in Table 1, then CS in (1) belongs to $\mathcal{L}$ family.

Proposition 3. *Let $E(S)$ in (1) depend on the marginal probabilities. If S is in $\mathcal{L}$ family, then CS in (1) is in $\mathcal{L}$ family.*

Proof: Expectation $E(S) = E[\lambda_1 + \mu_1(a+d)]$ is a function of the marginal probabilities. Thus $E(a+d)$, λ, and μ in (3) are functions of the marginal proportions. Equation (3) can therefore be written in a form $\lambda_2 + \mu_2(a+d)$ where

$$\lambda_2 = \frac{-E(a+d)}{\mu_1^{-1}(1-\lambda_1) - E(a+d)} \quad \text{and} \quad \mu_2 = \frac{1}{\mu_1^{-1}(1-\lambda_1) - E(a+d)}. \qquad \square$$

Examples of corrected coefficients that are in $\mathcal{L}$ family are S_{Scott}, S_{Cohen}, S_{GK3}, and S_{Ham}. These coefficients may be considered as corrected coefficients as well as ordinary coefficients that may be corrected for agreement due to chance. For example, S_{Scott}, S_{GK3}, and S_{Ham} (and S_{Cohen}) become S_{Cohen} after correction (1) if expectation $E(a+d)_{\text{Cohen}}$ is used. Coefficient S_{Mak} cannot be written in a form $\lambda + \mu(a+d)$, and does therefore not belong to $\mathcal{L}$ family.

At the end of this section we consider the following problem. Suppose a coefficient S in $\mathcal{L}$ family is corrected twice, using two different expectations, $E(a+d)$ and $E(a+d)^*$. Let the corrected coefficients be given by

$$CS = \frac{a+d-E(a+d)}{\mu^{-1}(1-\lambda) - E(a+d)} \quad \text{and} \quad CS^* = \frac{a+d-E(a+d)^*}{\mu^{-1}(1-\lambda) - E(a+d)^*}.$$

Note that $\mu^{-1}(1-\lambda)$ corresponding to coefficient S, is the same in both CS and CS^*. The problem is then as follows: if $E(a+d) \geq E(a+d)^*$, how are CS and CS^* related? It turns out that CS is a decreasing function of $E(a+d)$. Proposition 4 is limited to coefficients in $\mathcal{L}$ family of which the maximum value is 1, that is

$$\lambda + \mu(a+d) \leq 1 \quad \text{if and only if} \quad \frac{1-\lambda}{\mu} \geq (a+d).$$

It can be verified that the similarity coefficients in Table 2 and S_{Mak} satisfy this condition.

Proposition 4. *CS is a decreasing function of $E(a+d)$.*

Proof: $CS \leq CS^*$ if and only if

$$E(a+d)\left[\frac{1-\lambda}{\mu} - (a+d)\right] \geq E(a+d)^*\left[\frac{1-\lambda}{\mu} - (a+d)\right].$$

The requirement $\lambda + \mu(a+d) \leq 1$ completes the proof. $\square$

In the following, let $S = \lambda + \mu(a+d)$ be in $\mathcal{L}$ family and let

$$CS_{\text{Name}} = \frac{a+d-E(a+d)_{\text{Name}}}{\mu^{-1}(1-\lambda) - E(a+d)_{\text{Name}}}$$

be a corrected coefficient using expectation $E(a+d)_{\text{Name}}$. Using specific expectations $E(a+d)$ in combination with Proposition 4, we obtain the following result.

Theorem 3. *It holds that $CS_{\text{GK}} \overset{\text{(i)}}{\leq} CS_{\text{Scott}} \overset{\text{(ii)}}{\leq} CS_{\text{Cohen}}$.*

Proof: (i): Due to Proposition 4, it must be shown that $E(a+d)_{\text{GK}} \geq E(a+d)_{\text{Scott}}$. Suppose $(p_1+p_2) \geq (q_1+q_2)$. We have

$$
\begin{aligned}
&E(a+d)_{\text{GK}} \geq E(a+d)_{\text{Scott}},\\
&\frac{p_1+p_2}{2} \geq \left(\frac{p_1+p_2}{2}\right)^2 + \left(\frac{q_1+q_2}{2}\right)^2,\\
&\frac{p_1+p_2}{2}\left(1-\frac{p_1+p_2}{2}\right) \geq \left(\frac{q_1+q_2}{2}\right)^2,\\
&\frac{p_1+p_2}{2}\left(\frac{q_1+q_2}{2}\right) \geq \left(\frac{q_1+q_2}{2}\right)^2,\\
&(p_1+p_2) \geq (q_1+q_2).
\end{aligned}
$$

(ii): It must be shown that $E(a+d)_{\text{Scott}} \geq E(a+d)_{\text{Cohen}}$. We have

$$\left(\frac{p_1+p_2}{2}\right)^2 \geq p_1 p_2 \tag{14}$$

if and only if

$$\left(\frac{p_1-p_2}{2}\right)^2 \geq 0. \tag{15}$$

Furthermore, we have

$$\left(\frac{q_1+q_2}{2}\right)^2 \geq q_1 q_2 \tag{16}$$

if and only if

$$\left(\frac{q_1-q_2}{2}\right)^2 \geq 0. \tag{17}$$

Inequalities (14) and (16) are true because (15) and (17) are true. Adding (14) and (16) we obtain the desired inequality. □

Blackman and Koval (1993, p. 216) derived the inequality $S_{\text{Scott}} \leq S_{\text{Cohen}}$. Note that this inequality follows from the more general result Theorem 3 by using a coefficient S for which (4) holds (all coefficients in Table 2).

6. Multicategorical Generalization

Suppose the data consist of two nominal variables with identical categories, e.g. two psychologists each distribute m people among a set of k mutually exclusive categories. Let $\mathbf{N}$ be a contingency table with entries n_{ij}, where n_{ij} indicates the number of persons placed in category i by the first psychologist and in category j by the second psychologist. Furthermore, let n_{i+} and n_{+j} denote the marginal counts (row and column totals) of $\mathbf{N}$. Moreover, suppose that the categories of both variables are in the same order, so that the diagonal elements n_{ii} reflect the number of people put in the same category by the psychologists. If the variables are dichotomized, $m^{-1}\mathbf{N}$

equals Table 1. A straightforward measure of similarity is the observed proportion of agreement given by

$$P = \frac{1}{m}\sum_{i=1}^{k} n_{ii} = \frac{\text{tr}(\mathbf{N})}{m}.$$

Using $S = P$ in (1) we obtain

$$\frac{P - E(P)}{1 - E(P)}. \tag{18}$$

Goodman and Kruskal (1954), Scott (1955), and Cohen (1960) proposed measures that incorporate correction for chance agreement of a form (18). The different expectations $E(P)$ are defined as follows.

$$\text{No underlying continua:} \qquad E(P)_{\text{GK}} = \max_i^k\left(\frac{n_{i+} + n_{+i}}{2m}\right).$$

$$\text{One frequency distribution:} \qquad E(P)_{\text{Scott}} = \sum_{i=1}^{k}\left(\frac{n_{i+} + n_{+i}}{2m}\right)^2.$$

$$\text{Two frequency distributions:} \qquad E(P)_{\text{Cohen}} = \frac{1}{m^2}\sum_{i=1}^{k} n_{i+}n_{+i}.$$

Note that P is a natural extension of $S_{\text{SM}} = a + d$ to nominal variables. Family $\mathcal{L}$ can be extended to coefficients of a form $S = \lambda + \mu P$, where λ and μ, unique for each coefficient, depend on the marginal probabilities of contingency table $\mathbf{N}$. All results for the 2×2 case naturally generalize to coefficients of a form $S = \lambda + \mu P$. Coefficient P and the multicategorical versions of S_{GK3}, S_{Scott}, and S_{Cohen} that are obtained by using expectations $E(P)_{\text{GK}}$, $E(P)_{\text{Scott}}$, and $E(P)_{\text{Cohen}}$ in (18), belong to $\mathcal{L}$ family (have a form $S = \lambda + \mu P$; note Proposition 3). Furthermore, it is not difficult to show that ratio (4) holds for multicategorical coefficients P, S_{GK3}, S_{Scott}, and S_{Cohen}. In this section only the generalization of Proposition 1, the powerful result by Albatineh et al. (2006), is presented.

Proposition 1b. *Two members in $\mathcal{L}$ family become identical after correction* (1) *if they have the same ratio $\mu^{-1}(1 - \lambda)$.*

Proof: $E(S) = E(\lambda + \mu P) = \lambda + \mu E(P)$ and consequently the corrected coefficient becomes

$$CS = \frac{P - E(P)}{\mu^{-1}(1 - \lambda) - E(P)}.$$

□

7. Multivariate Generalization

Multivariate coefficients may be used to determine the degree of agreement among three or more raters in psychological assessment, or to compare partitions from three different cluster algorithms. Multivariate versions of Cohen's kappa (S_{Cohen}) can for instance be found in Fleiss (1971), Light (1971), Popping (1983), and Heuvelmans and Sanders (1993).

Suppose we want to determine the agreement among k raters. Similar to Table 1, we may construct $k(k-1)/2$ bivariate 2×2 tables: each proportion table compares two variables i and j.

Let a_{ij} denote the proportion of people that possess a characteristic according to both psychologists i and j, let d_{ij} denote the proportion of people that lack the characteristic according to both psychologists, and let p_i denote the proportion of people that possess the characteristic according to psychologist i. Family $\mathcal{L}$ may be extended to a multivariate family $\mathcal{L}^{(k)}$ of coefficients of a form

$$\lambda^{(k)} + \frac{2\mu^{(k)}}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (a_{ij} + d_{ij}),$$

where $\lambda^{(k)}$ and $\mu^{(k)}$ depend on the marginal probabilities of the 2×2 tables only. Note that

$$S_{\text{SM}}^{(k)} = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (a_{ij} + d_{ij})$$

is a straightforward multivariate generalization of S_{SM}. Quantity $2/k(k-1)$ is used to ensure that the value of coefficient $S_{\text{SM}}^{(k)}$ lies between 0 and 1. Let us present some other examples of coefficients that belong to $\mathcal{L}^{(k)}$ family.

Example 1b. A three-way formulation of $S_{\text{Cze}} = 2a_{12}/(p_1 + p_2)$ (Example 1), such that the coefficient is a linear transformation of $S_{\text{SM}}^{(3)}$, is given by

$$S_{\text{Cze}}^{(3)} = \frac{a_{12} + a_{13} + a_{23}}{p_1 + p_2 + p_3} = \frac{3S_{\text{SM}}^{(3)} - 3}{2(p_1 + p_2 + p_3)} + 1.$$

A general multivariate version of S_{Cze} is given by

$$S_{\text{Cze}}^{(k)} = \frac{2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij}}{(k-1)\sum_{i=1}^{k} p_i} = \frac{k[S_{\text{SM}}^{(k)} - 1]}{2\sum_{i=1}^{k} p_i} + 1.$$

Coefficient $S_{\text{Cze}}^{(k)}$ can be written in a form $S_{\text{Cze}}^{(k)} = \lambda^{(k)} + \mu^{(k)} S_{\text{SM}}^{(k)}$, where

$$\lambda^{(k)} = \frac{-k}{2\sum_{i=1}^{k} p_i} + 1 = 1 - \mu^{(k)} \quad \text{and} \quad \mu^{(k)} = \frac{k}{2\sum_{i=1}^{k} p_i}.$$

Quantities $\lambda^{(k)}$ and $\mu^{(k)}$ naturally extend λ and μ corresponding to S_{Cze} in Example 1.

Example 3b. Popping (1983) and Heuvelmans and Sanders (1993) describe the same multivariate extension of Cohen's (1960) kappa. For k dichotomized variables, the multivariate kappa is given by

$$S_{\text{Cohen}}^{(k)} = \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}(a_{ij} + d_{ij} - p_i p_j - q_i q_j)}{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}(p_i q_j + p_j q_i)}.$$

Coefficient $S_{\text{Cohen}}^{(k)}$ can be written in a form $S_{\text{Cohen}}^{(k)} = \lambda^{(k)} + \mu^{(k)} S_{\text{SM}}^{(k)}$, where

$$\lambda^{(k)} = \frac{-\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}(p_i p_j + q_i q_j)}{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}(p_i q_j + p_j q_i)} \quad \text{and} \quad \mu^{(k)} = \frac{k(k-1)}{2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}(p_i q_j + p_j q_i)}.$$

Using the heuristics in Examples 1b and 3b one may obtain multivariate formulations of other coefficients in Table 2. The remainder of the section is used to present generalizations of Proposition 1, the main result in Albatineh et al. (2006), and Corollary 1. Both extensions show that family $\mathcal{L}^{(k)}$ naturally generalizes family $\mathcal{L}$, with respect to correction (1), to multivariate coefficients.

Proposition 1c. *Two members in $\mathcal{L}^{(k)}$ family become identical after correction (1) if they have the same ratio*

$$\frac{1-\lambda^{(k)}}{\mu^{(k)}}. \tag{19}$$

Proof:

$$E\big[S^{(k)}\big] = \lambda^{(k)} + \mu^{(k)} E\left[\frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (a_{ij} + d_{ij})\right] = \lambda^{(k)} + \mu^{(k)} E\big[S_{\mathrm{SM}}^{(k)}\big].$$

Consequently, the corrected coefficient becomes

$$CS^{(k)} = \frac{S_{\mathrm{SM}}^{(k)} - E[S_{\mathrm{SM}}^{(k)}]}{(1-\lambda^{(k)})/\mu^{(k)} - E[S_{\mathrm{SM}}^{(k)}]}.$$

□

Corollary 1b. *Coefficients $S_{\mathrm{SM}}^{(k)}$, $S_{\mathrm{Cze}}^{(k)}$, and $S_{\mathrm{Cohen}}^{(k)}$ become equivalent after correction (1).*

Proof: Using the formulas of $\lambda^{(k)}$ and $\mu^{(k)}$ corresponding to each coefficient, we obtain the ratio (19)

$$\frac{1-\lambda^{(k)}}{\mu^{(k)}} = 1$$

for all three coefficients. Obtaining ratio (19) for coefficients $S_{\mathrm{SM}}^{(k)}$ and $S_{\mathrm{Cze}}^{(k)}$ is straightforward. Using the formulas for $\lambda^{(k)}$ and $\mu^{(k)}$ from Example 3b, we obtain the ratio (19)

$$\frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (p_i q_j + p_j q_i + p_i p_j + q_i q_j) = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (p_i + q_i)(p_j + q_j) = 1.$$

□

8. Discussion

The inspiration for this work came from the paper by Albatineh et al. (2006), who studied correction for chance for similarity coefficients from a general perspective. For a specific family of coefficients they showed that coefficients may coincide after correction for chance, irrespective of what expectation is used.

The study of correction for chance presented in this paper focused on resemblance measures for 2×2 tables. It is surprising how much output has been generated for this simple case (Pearson, 1947; Fleiss, 1975; Gower and Legendre, 1986; Krippendorff, 1987; Mak, 1988; Blackman and Koval, 1993; Albatineh et al., 2006; Warrens, 2008, in press). Furthermore, for the 2×2 case we have many similarity coefficients at our disposal, and some of these were used to illustrate the results in this paper. As suggested by the multicategorical and multivariate generalizations in

Sections 6 and 7, the properties derived in this paper apply to coefficients of a form $S = \lambda + \mu x$, for which we have

$$E(S) = E[\lambda + \mu x] = \lambda + \mu E(x), \tag{20}$$

where λ and μ depend on the marginals of the table corresponding to the data type. Property (20) is central in Proposition 1, the main result in Albatineh et al. (2006), and several other results in this paper. The general coefficients for metric scales in Zegers and Ten Berge (1985), for instance, satisfy condition (20).

References

Albatineh, A.N., Niewiadomska-Bugaj, M., & Mihalko, D. (2006). On similarity indices and correction for chance agreement. *Journal of Classification*, *23*, 301–313.

Baulieu, F.B. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, *6*, 233–246.

Blackman, N.J.-M., & Koval, J.J. (1993). Estimating rater agreement in 2 × 2 tables: Correction for chance and intraclass correlation. *Applied Psychological Measurement*, *17*, 211–223.

Bloch, D.A., & Kraemer, H.C. (1989). 2 × 2 Kappa coefficients: Measures of agreement or association. *Biometrics*, *45*, 269–287.

Bray, J.R. (1956). A study of mutual occurrence of plant species. *Ecology*, *37*, 21–28.

Brennan, R.L., & Light, R.J. (1974). Measuring agreement when two observers classify people into categories not defined in advance. *British Journal of Mathematical and Statistical Psychology*, *27*, 154–163.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Czekanowski, J. (1932). Coefficient of racial likeliness und Durchschnittliche Differenz. *Anothropologidcher*, *14*, 227–249.

Dice, L.R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*, 297–302.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382.

Fleiss, J.L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, *31*, 651–659.

Gleason, H.A. (1920). Some applications of the quadrat method. *Bulletin of the Torrey Botanical Club*, *47*, 21–33.

Goodman, L.A., & Kruskal, W.H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*, 732–764.

Gower, J.C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, *3*, 5–48.

Hamann, U. (1961). Merkmalsbestand und Verwandtschaftsbeziehungen der Farinose. Ein Betrag zum System der Monokotyledonen. *Willdenowia*, *2*, 639–768.

Heuvelmans, A.P.J.M., & Sanders, P.F. (1993). Beoordelaarsovereenstemming. In T.J.H.M. Eggen & P.F. Sanders (Eds.), *Psychometrie in de praktijk* (pp. 443–470). Arnhem: Cito Instituut voor Toetsontwikkeling.

Hubálek, Z. (1982). Coefficients of association and similarity based on binary (presence-absence) data: An evaluation. *Biological Reviews*, *57*, 669–689.

Hubert, L.J. (1977). Nominal scale response agreement as a generalized correlation. *British Journal of Mathematical and Statistical Psychology*, *30*, 98–103.

Hubert, L.J., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Jaccard, P. (1912). The distribution of the flora in the Alpine zone. *The New Phytologist*, *11*, 37–50.

Krippendorff, K. (1987). Association, agreement, and equity. *Quality and Quantity*, *21*, 109–123.

Light, R.J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *76*, 365–377.

Mak, T.K. (1988). Analyzing intraclass correlation for dichotomous variables. *Applied Statistics*, *37*, 344–352.

Morey, L.C., & Agresti, A. (1984). The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement*, *44*, 33–37.

Nei, M., & Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, *76*, 5269–5273.

Pearson, E.S. (1947). The choice of statistical tests illustrated on the interpretation of data classed in a 2 × 2 table. *Biometrika*, *34*, 139–167.

Popping, R. (1983). *Overeenstemmingsmaten voor nominale data*. Ph.D. thesis, Groningen, Rijksuniversiteit Groningen.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*, 846–850.

Rogot, E., & Goldberg, I.D. (1966). A proposed index for measuring agreement in test-retest studies. *Journal of Chronic Disease*, *19*, 991–1006.

Scott, W.A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, *19*, 321–325.

Sokal, R.R., & Michener, C.D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, *38*, 1409–1438.

Sokal, R.R., & Sneath, P.H. (1963). *Principles of numerical taxonomy*. San Francisco: Freeman.

Sørenson, T. (1948). A method of stabilizing groups of equivalent amplitude in plant sociology based on the similarity of species content and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab Biologiske Skrifter*, *5*, 1–34.

Steinley, D. (2004). Properties of the Hubert–Arabie adjusted Rand index. *Psychological Methods*, *9*, 386–396.

Warrens, M.J. (2008, in press). On the indeterminacy of resemblance measures for binary (presence/absence) data. *Journal of Classification*.

Zegers, F.E. (1986). *A General family of association coefficients*. Ph.D. thesis, Groningen, Rijksuniversiteit Groningen.

Zegers, F.E., & Ten Berge, J.M.F. (1985). A family of association coefficients for metric scales. *Psychometrika*, *50*, 17–24.

Manuscript received 16 FEB 2007
Final version received 21 DEC 2007
Published Online Date: 1 MAR 2008